

Amendments to the Claims:

This listing of claims will replace all prior versions, and listings of claims in the application:

Listing of Claims:

1 1. (Currently Amended) A computer implemented method of identifying and
2 extracting content from HTML formatted web pages, comprising the steps of:

3 selecting a model page, wherein the model page includes a plurality of HTML
4 tags;

5 identifying a first area of interest in the model page;

6 parsing the model page to ~~determine~~ generate a first string of symbols associated
7 with the plurality of HTML tags, wherein the first area of interest is identified by a first portion
8 of the first string of symbols;

9 retrieving a second web page associated with a different URL than the model
10 page;

11 parsing the second web page to ~~determine~~ generate a second string of symbols
12 associated with the HTML tags of the second web page; and

13 comparing the first and second strings to determine whether the second string
14 includes a second portion similar to the first portion of the first string, wherein the second
15 portion corresponds to a second area of interest in the second page.

1 2. (Original) The method of claim 1, wherein the step of comparing
2 includes applying an approximate pattern matching algorithm to the first and second strings.

1 3. (Original) The method of claim 1, further comprising the step of
2 storing the first and second areas of interest in a database.

1 4. (Original) The method of claim 1, further comprising the step of
2 extracting the second area of interest from the second page.

1 5. (Original) The method of claim 4, further comprising the step of
2 applying a regular expression matching algorithm to the extracted second area of interest.

1 6. (Original) The method of claim 1, wherein the first and second areas
2 of interest each include two or more distinct sub-areas of the respective page.

1 7. (Original) The method of claim 1, wherein the step of identifying a
2 first area of interest includes the step of identifying portions of the HTML tags of the model
3 page.

1 8. (Original) The method of claim 1, wherein the step of identifying a
2 first area of interest is performed using a manual pointing and selecting device.

1 9. (Original) The method of claim 1, wherein the steps of selecting and
2 identifying are performed manually and wherein the remaining steps are performed
3 automatically.

1 10. (Original) The method of claim 1, wherein the second web page is
2 retrieved from a remote website over the Internet.

1 11. (Original) The method of claim 1, wherein the HTML tags include
2 attributes and attribute values.

1 12. (Currently amended) A computer readable medium containing
2 instructions for controlling a computer system to automatically identify and extract desired
3 content from a retrieved HTML formatted web page, by automatically:
4 parsing the HTML code of a manually selected model web page to ~~determine~~
5 generate a first string of symbols associated with a first plurality of HTML tags;

6 retrieving a second web page associated with a different URL than the model web
7 page;
8 parsing the HTML code of the second web page to ~~determine~~ generate a second
9 string of symbols associated with HTML tags of the second page; and
10 comparing the first and second strings to determine whether the second page
11 includes a second plurality of HTML tags substantially matching the first plurality of HTML
12 tags.

1 13. (Original) The computer readable medium of claim 12, wherein the
2 first plurality of HTML tags are identified by an operator using a pointing and selection device
3 coupled to the computer system.

1 14. (Original) The computer readable medium of claim 12, wherein the
2 second web page is retrieved from a remote website over the Internet.

1 15. (Original) The computer readable medium of claim 12, further
2 including instructions for extracting a portion of the second page corresponding to the second
3 plurality of HTML tags.

1 16. (Original) The computer readable medium of claim 15, wherein the
2 instructions further control the computer system to store the extracted portion of the second page
3 in a database.

1 17. (Original) The computer readable medium of claim 15, further
2 including instructions for controlling the computer system to apply a regular expression
3 matching algorithm to the extracted portion of the second page.

1 18. (Original) The computer readable medium of claim 15, wherein the
2 extracted portion of the second page includes two or more distinct sub-areas.

1 19. (Original) The computer readable medium of claim 12, wherein the
2 instructions for comparing include instructions for applying an approximate string matching
3 algorithm to the first and second strings.

1 20. (Original) The computer readable medium of claim 12, wherein the
2 HTML tags include attributes and attribute values.

1 21. (Currently amended) A computer system for identifying and extracting
2 content from HTML formatted web pages, the system comprising:
3 means for retrieving web pages including HTML tags, wherein a model web page
4 is retrieved;
5 means for manually identifying a first area of interest in the model page, wherein
6 the first area of interest corresponds to a first plurality of HTML tags; and
7 a processor including:
8 means for parsing a page, wherein the parsing means parses the model
9 page ~~to determine~~ and generates a first string of symbols associated with the first plurality of
10 HTML tags, and wherein the parsing means thereafter parses an automatically retrieved second
11 web page associated with a different URL than the model page and generates ~~to determine~~ a
12 second string of symbols associated with the HTML tags of the second web page;
13 means for comparing the first and second strings to determine whether the
14 second string includes a second portion similar to the first portion of the first string, wherein the
15 second portion corresponds to a second area of interest in the second page; and
16 means for extracting the second area of interest from the second page.

1 22. (Currently amended) A computer implemented method of identifying and
2 extracting content from web pages formatted using a markup language, comprising the steps of:
3 selecting a model page, wherein the model page includes a plurality of tokens;
4 identifying a first area of interest in the model page;

5 parsing the model page to ~~determine~~ generate a first string of symbols associated
6 with the plurality of tokens, wherein the first area of interest is identified by a first portion of the
7 first string of symbols;

8 retrieving a second web page associated with a different URL than the model
9 page;

10 parsing the second web page to ~~determine~~ generate a second string of symbols
11 associated with the tokens of the second web page; and

12 comparing the first and second strings to determine whether the second string
13 includes a second portion similar to the first portion of the first string, wherein the second
14 portion corresponds to a second area of interest in the second page.

1 23. (Original) The method of claim 22, further comprising the step of
2 extracting the second area of interest from the second page.

1 24. (Original) The method of claim 22, wherein the markup language is
2 selected from the group consisting of HTML, XML, WML, DHTML and HDML.

1 25. (Original) The method of claim 22, wherein the tokens include tag
2 elements and text elements.

1 26. (New) A computer-implemented method of identifying similar content in
2 HTML formatted web pages, the method comprising:
3 selecting a model page, wherein the model page includes a plurality of HTML
4 tags;
5 identifying a first area of interest in the model page;
6 generating a first string of symbols for the plurality of HTML tags associated with
7 the first area of interest;
8 retrieving a second web page associated with a different URL than the model
9 page;

10 generating a second string of symbols for the HTML tags of the second web page;
11 and
12 comparing the first and second strings to determine whether the second string
13 includes a portion similar to the first string, wherein the portion corresponds to a second area of
14 interest in the second page.

1 27. (New) The method of claim 26, further comprising extracting the second
2 area of interest from the second page.

1 28. (New) The method of claim 26, wherein identifying is performed
2 manually using a user-input device.